Statistical Population Dataset version 4: Research to Date and Future Developments (MARP)

Table of Contents

Executive Summary 3
MARP Ask 4
Evolution of the Statistical Population Dataset 5
How is an SPD made? 5
SPD v1 5
SPD v2 6
SPD v3 6
SPD v4 7
Department for Work and Pensions data 8
SPD v4.0 'starting point' 8
New Data Sources Research 9
Hospital Episode Statistics and Emergency Care Dataset 9
Individualised Learner Record 10
Deaths Registrations 10
Inclusion Rules Research 11
Presence-based PDS rules 11
Quality Research 12
Analysis of SPD v4.1 12
Development of SPD v4.2 13
Census/CCS to DI linkage 14
Comparing the coverage of the SPD, DI and Census 16
Future Development and Next Steps 17
Delivering our "best-possible" SPD version 4 for the Recommendation 17
Longer term developments 17
Modelled inclusion rules 17
Estimation 18
Determining when to stop/start SPD development in future 18
Conclusion 19
References 20
Annex 22
Annex 1: Data cleaning and pre-processing applied to data sources in SPD version 4.1 22
Annex 2: Comparison of deaths registration data to SPD v2 and CIS 23

Comparisons to SPD v2 23
Comparisons to CIS 23

Executive Summary

This October 2022 update adds further detail to the development of SPD v4, and includes a new section exploring how we might develop criteria by which to determine whether/when SPD development should stop/start in future.

Population and migration statistics underpin a wide variety of other statistics and are vital to support decisions about local services and inform public debate. The current population statistics system relies on the decennial Census and, as we move further away from Census Day, official estimates suffer from increasing bias and uncertainty. Official estimates for the mid-year population are also currently produced on an annual basis the following June.

We identified a need for a more timely and accurate understanding of the population for our users to meet the biggest challenges facing society. We are therefore transforming the way we produce population and migration statistics to make better use of administrative sources to produce the best statistics from the best-available data at any given point in time. Our essential user needs for a future population and migration statistics system were reviewed by the Methodological Assurance Review Panel in March 2021 (Bartleet, 2021).

In 2023, the National Statistician will report to the government on our progress, including the transformation we have already delivered, and set out what is needed in the future to continue to achieve these ambitions.

ONS is developing a Dynamic Population Model (DPM) (Blackwell, 2022) to address the needs for timely sub-national population estimates as well as coherence between population stocks and flows in a transformed system. The DPM will use a statistical modelling approach to draw strength from a range of data sources and demographic insights, such as administrative and survey data, to produce monthly and annual estimates of the population. One input the DPM is using as it develops are the Statistical Population Datasets.

The Statistical Population Dataset (SPD) is a mid-year approximation of the usually resident population of England and Wales using administrative data records. We have been producing SPDs for several years and have iteratively developed our method in response to quality evaluations undertaken against official population estimates. Administrative data sources are not collected for the purpose of producing population statistics and can vary in their coverage, collection processes and variable definitions. Our mission is to combine a range of sources into one consistent picture of the population, which has high-quality coverage by age, sex and Local Authority[1] (Office for National Statistics, 2013) and accurately captures flows into and out of the population.

We have previously published three methods for producing Statistical Population Datasets, SPD v1 (Office for National Statistics, 2015), SPD v2 (Office for National Statistics, 2016) and SPD v3 (Office for National Statistics, 2019). The mid-2020 version 3 estimates compared very favourably to the 2021 Census estimates and were an important part of the evidence packs provided to local authorities during the Census quality assurance process.

Since December 2021, our focus has been to build on our learnings from evaluating v2 and v3 and develop a new version of the SPD (version 4) using a greater selection of sources that have become available to us. Our researchers did not have access to Department for Work and Pensions (DWP)

[1] The Beyond 2011 Options Report states that to meet the "maximum quality achieved in the current system" is all LA estimates are unbiased and have a 95% confidence interval of +/- 3.8% or better. With the development of the DPM, we will need to think more broadly about our range of quality measures.

datasets for a period from January 2022, which constrained early development of SPD version 4 as these sources have good coverage for the working age and pensionable age groups.

Our approach for the first iteration of version 4 (SPD v4.1) was to produce the best approximation of the usually resident population from the remaining administrative data available to us and show what is possible without DWP datasets. So far, development has researched how potential new data sources can improve the quality of the SPD; how the inclusion rules can be improved; and how to integrate quality considerations into our production process.

SPD version 4 is still in development, both SPD v3 (mid-2020) and SPD v4 (mid-2021) will be tested through linkage to Census 2021 to evaluate the strengths and limitations of each method. We also plan to re-introduce the DWP data sources to the SPD when they become available, which will provide additional evidence to evaluate assumptions in the v4.1 method against income-related activity.

We plan to use the insight we gain in this process to produce a second, "best-possible" SPD v4 iteration (SPD v4.2), which includes DWP datasets, by December 2022. By evaluating the under- and over-coverage of SPD v4.2 through Census 2021 linkage and reproducing associated uncertainty intervals (Blackwell, 2020), we will be able to explore the estimation challenge further. Our aim is to produce our 'best' quality SPD and evaluate and communicate the quality of the outputs using these 2021 Census comparisons. Both the SPDs and the quality information will provide evidence for the 2023 NS Recommendation and importantly, feed into the development of the DPM as we move towards the transformed population and social statistics system. To ensure the robustness of our approach to the SPD, and transformed system, we are seeking an expert panel review of our progress to date and future plans.

MARP Ask

- Oct update – to see progress made with DWP data and to provide comment on the criteria for when to start/stop SPD development in the future.
- To endorse our plans to further develop SPD v4 during 2022
- To endorse the need for tax and benefits activity analysis to feed into the SPD v4 method
- To provide any additional comments on the proposed SPD development that could improve them as an input to the DPM.

Evolution of the Statistical Population Dataset

The Statistical Population Dataset (SPD)[2] approximates the mid-year usually resident population[3] stock of England and Wales using administrative data records. We aim to produce the highest quality SPDs by age, sex and local authority at a given mid-year reference point[4] as soon as possible after that point. While our administrative data sources can span a range of periods, our inclusion rules are designed to align these supplies as closely as possible with the reference point and "usually resident" definition.

We have previously published three versions of SPD methods: version 1 (Office for National Statistics, 2015), version 2 (Office for National Statistics, 2016) and version 3 (Office for National Statistics, 2019). Version 2 and version 3 were both used as evidence in the quality assurance process for Census 2021 estimates and were included in the figures delivered as part of the Local Authority engagement exercise. While the Dynamic Population Model was in its early stages of development in 2022, time-series of SPD v2 and SPD v3 have been used as population stock inputs (Blackwell, et al., 2021).

How is an SPD made?

1. Acquire the admin data sources and bring them together in the same environment.
2. A dedicated team constructs the Demographic Index (DI) by linking and de-identifying admin data sources to create a single entity and ID (cluster of records) for what we believe is a single individual across multiple different sources[5]. We can also link data sources which are not built into the DI if they share a source-specific identifier (e.g., NHS number) with a source that is.
3. Collapse each cluster of DI records into a single row, creating an SPD 'spine' with a list of data source IDs for each person across all the sources we are interested in. Undertake data-cleaning and pre-processing to filter out incomplete records.[6]
4. Filter the spine by a set of researched inclusion rules, with the aim to retain those individuals who our rules suggest are usually resident and exclude those that are not.
5. Assign an age, sex and local authority/output area for each record based on a set of rules to derive this from their admin data records.
6. Count records by age, sex and local authority to produce aggregated population counts by LA, age and sex.

SPD v1

SPD v1 (Office for National Statistics, 2015) linked administrative records between three datasets:

- Patient Register (PR) – *National Health Service*
- Customer Information System (CIS) – *Department for Work & Pensions*

[2] Note: Between June 2019 and December 2021, ONS used Admin-based Population Estimate (ABPE) to denote its population stock of admin data records. With the development of the DPM as a coherent modelling system, we returned to calling the population stock the Statistical Population Dataset (SPD), and the modelled estimates ABPEs.

[3] We are currently adopting the UN (2008) definition of "usually resident". That is, the place at which a person has lived continuously for at least 12 months, not including temporary absences, or intends to live at for at least 12 months.

[4] The mid-year SPD reference point is 30th June.

[5] The work to develop the Demographic Index has been reviewed by the Longitudinal Study Advisory Panel in October 2021 and will be reviewed by the Methods and Research Assurance Group in July 2022.

[6] A summary of pre-processing steps can be found in Annex 1

- Higher Education Statistics Agency (HESA)

SPD v1 had some under-coverage in school-aged children and the method was soon developed into SPD version 2 by adding School Censuses. At this point we focused solely on evaluating and improving v2.

SPD v2

The SPD v2 (previously called ABPE v2) method was first published in November 2016 (Office for National Statistics, 2016). The single sources used in SPD v2 were:

- Patient Register (PR) – *National Health Service*
- Customer Information System (CIS) – *Department for Work & Pensions*
- Higher Education Statistics Agency (HESA)
- English School Census (ESC) – *Department for Education*
- Welsh School Census (WSC) – *Welsh Government*

A date of death is entered on the DWP CIS record after someone dies, and this is used to exclude such records from the SPD. Deaths can be derived from CIS to a high degree of accuracy (see Annex 2), however lags in death reporting may mean that some people that have died before the reference date are included in the SPD (Blake, 2020).

The inclusion rules for v2 were **presence-based**, meaning a record is included if a successful link is found between records on at least two sources[7]. Including the correct people in SPD v2 is dependent on making correct links between sources, and the method is vulnerable to inconsistencies if linkage error is present.

Previous research shows that SPD v2 typically over-estimates the population compared with official estimates, particularly for males aged 30-54 (Office for National Statistics, 2017).

When we analysed SPD v2 over time, we identified that cohorts of people born in the same year grew year-on-year (Blake, 2021). While we would expect some net migration in the younger working age population (aged 20 to 40 years), the levels of list inflation in SPD v2 were much higher than in the MYEs. National Insurance number records are not removed from DWP CIS when the individual emigrates, so the method was vulnerable to not excluding these individuals if they remained present on other sources.

SPD v3

We published the SPD v3 (previously ABPE v3) method in June 2019 (Office for National Statistics, 2019). The design incorporated more data sources into the methodology:

- Patient Register (PR) – *National Health Service*
- Patient Demographics Service (PDS) – *National Health Service*
- Customer Information System (CIS) – *Department for Work and Pensions*
- Higher Education Statistics Agency (HESA)
- English School Census (ESC) – *Department for Education*
- Welsh School Census (WSC) – *Welsh Government*
- Benefits and Income Dataset (BIDS) – *Department for Work and Pensions*
- Births

[7] For those aged under 4-years-old, presence on only PR is required for inclusion in SPD v2.

As with SPD v2, the date of death record in CIS was used to remove people who have died from SPD v3.

SPD v3 used **activity-based** inclusion rules to attempt to address over-coverage patterns in SPD v2. An individual is assumed part of the usually resident population if they have interacted with one or more data sources in the 12 months prior to the mid-year reference point and are not flagged as being resident outside of England and Wales.[8].

Evaluation against official estimates found that it generally returned lower estimates, particularly for working age populations up to 65 years (Blake, 2021). While some over-coverage remained, the age cohorts over time seem fairly stable, suggesting the activity-based rules are more likely to remove people who have left England and Wales. The mid-2020 SPD v3 was a particularly useful resource in the Census 2021 quality assurance process and was found to broadly track the provisional estimates, which has increased confidence in our ability to approximate the usually resident population using admin-based methods.

We did however find some instability in SPD v3 for those age groups where people are at transition points in their lives, notably around younger adult ages.

If an individual appears across multiple sources in SPD v3, their age, sex and geography attributes are assigned according to the following hierarchy of data sources:

- HESA > Births > PDS > SC > BIDS/CIS

This system was informed by our understanding of the quality of each data source at the time. With new data sources being added to future versions of the SPD, we plan to revisit this approach as we learn more about the quality of each source. We plan to use the results of the DI to 2021 Census linkage to inform this.

SPD v4

We signalled completion of work on SPD v2 and SPD v3 with a publication in November 2021, which introduced a time-series for both and highlighted the strengths and limitations of each version (Blake, 2021). In particular, the activity-based SPD v3 method had been designed towards under-coverage to work with a dual-system estimation approach. Despite these efforts, the net under-coverage of version 3 hid areas of over-coverage at smaller population levels and it proved extremely challenging to entirely eradicate over-coverage without a severe cost to quality.

With development of the Dynamic Population Model, new thinking about how the SPD best fitted into that was needed. At the same time, access to DWP Customer Information System data had to be withdrawn for a period, and the coverage gaps this produced in our starting point for version 4 made it not fit for use with the DPM. As a consequence, the development of SPD v4 started without access to DWP CIS, but our aim was to take all the elements of learning from SPD v2 and SPD v3 along with further developments to those methods. SPDv4 development is intended to address several areas.

We are researching whether and how **new data sources** can improve the quality of the SPD. We are exploring which sources have the greatest potential to improve the coverage, accuracy and

[8] It is also sufficient to be included in SPD v3 if a record has not interacted with a data source themselves but are related to and live with somebody that has (the “inactive relatives” rule).

inclusivity of the SPD. We investigate the quality, completeness and suitability of several new data sources and decide which to link to the SPD. We then measure each sources' value in terms of the additional records, recent activity or up-to-date age, sex and address information each provides.

We are researching how to optimise the **inclusion rules** that are applied to different admin sources in the method to retain members of the usually resident population, while excluding those that aren't. This involves understanding how individuals interact with the data sources over time to improve the rules for certain ages where people are likely to move from one administrative source to another e.g. School Census to HESA or HESA to PAYE.

We are integrating **quality** throughout the SPD production process, to ensure our outputs are reproducible and that we have the relevant quality metrics at each point in the process, including understanding the quality of the individual sources that feed into the SPD.

SPD v4 development is further supported by several cross-cutting research areas across the office. The **Demographic Index (DI)** is the underlying index of individuals on admin data from which the SPD is derived and ONS works to implement improvements and release new versions incrementally. **Linkage between Census/Census Coverage Survey and the DI** will help us quantify record-level over-/under-coverage of SPD v3 and different iterations of SPD v4 against Census 2021 and allow us to evaluate which sources provide the most up to date address information.

Department for Work and Pensions data

As noted above, our researchers lost access to DWP Customer Information System data in January 2022. This has constrained the initial phase of work on SPD v4.

So far in the SPD development journey, DWP and HMRC admin sources have been central to our methods. CIS was one of only 4 sources in the SPD v2 method and was responsible for contributing the majority of working age individuals. In SPD v3 the tax and benefits data (BIDS) played a central role. This contains a variety of benefits and income information from: Pay As You Earn, National Benefits Database, Single Housing Benefit Extract, Tax Credits, Personal Independence Payment, Child Benefit and Universal Credit.

The dependence on benefits and income data for activity in SPD v3 is exposed when we remove these sources from the v3 method - the 2020 SPD v3 loses around 38 million records from its overall England and Wales counts.

Despite the obvious challenge the lack of DWP data presented, we developed SPD v4.1 without income datasets, with the intention of reincorporating these sources when they became available to us. Investigating alternative methods and new sources to use as contingency would in turn increase the resilience of our SPD method to similar supply issues in the long-term. It also gave an opportunity to demonstrate the value of income datasets to population statistics and therefore strengthen the case for their ongoing supply.

We acquired access to a new supply of DWP and HMRC CIS and BIDs datasets in May 2022. Due to DWP system updates and the transferral of supply of certain BIDs data (PAYE), these datasets were treated as new data sources rather than a continuation of the previous supply.

SPD v4.0 'starting point'

Our starting point in SPD v4 development (SPD v4.0) was consequently a "non-DWP" version of SPD v3 along with some small changes. That is, all data sources used in SPD v3, under the same, activity-based inclusion rules, except for the income datasets that we no longer had access to. Additionally,

following prior analysis of SPD v3 and the single sources, we also removed the Patient Register and Births datasets, as the information we need from both can be obtained from the PDS. The data sources in our starting point included were therefore:

- Patient Demographics Service (PDS) – *National Health Service*
- Higher Education Statistics Agency (HESA)
- English School Census (ESC) – *Department for Education*
- Welsh School Census (WSC) – *Welsh Government*

SPD version 3 had been constructed towards under-coverage and a dual system estimation approach. When DWP data sets were removed, the SPD v4.0 'starting point' version produced extremely high under-coverage when compared to version 3 and the official estimates and was not fit for use in the DPM. It was necessary to readjust our approach towards strictly activity-based inclusion rules as well as explore the potential improvements from new data sources.

New Data Sources Research

New admin sources may have the potential to improve the coverage or accuracy of individuals in our SPDs and help target specific population groups. Including the wrong data, or incorporating sources in an improper way, may however also create over-coverage and linkage error in our SPDs and assign individuals the wrong attributes. So, it is important that we build a strong case for each new source through evidence and link them in a reliable way. Our start-to-finish process for researching a new data source for the SPD includes:

- Identifying which admin sources are/will be available for analysis in ONS, including ensuring that there exists data which covers our desired time-series of reference points (2016-2021) and considering the timeliness and sustainability of the ongoing supply of data.
- Investigating whether the variables in the data source contain the demographic information we would need on individuals and/or any other "signs of life".
- Investigating whether it is feasible to link the source to the current SPD, DI or single sources by a reliable identifier to allow us to assess the value of using the source.
- Making quality evaluations of the source, including missingness and basic counts. Linking the data source to the SPD to measure the value of including each source in terms of improving coverage, providing activity indicators or up-to-date demographic information.

Hospital Episode Statistics and Emergency Care Dataset

The PDS is a large dataset with between 75m-80m records and the SPD method uses this to add a large number of individuals to the SPD "spine". However, in activity-based rules, interactions with PDS are limited to little more than births/new registrations and name/address changes. These are not interactions which would generally occur on an annual basis for most of the usually resident population. So, the PDS is very limited as a source from which resident activity can be inferred.

Hospital Episode Statistics (HES) contains all interactions with Accident and Emergency, Outpatient or Admitted Care services at NHS hospitals in England. It does not contain any medical details. These tables are inclusive datasets and cover males and females across a wide age range. HES is initially provided at episode-level, usually with multiple rows per individual. Our pre-processing of HES (Annex 1) includes removing individuals with death flags and missing NHS numbers, before consolidating each individual's HES information into one row per person.

Our research found that the overwhelming majority of HES individuals can be found on PDS by their NHS number. Around 21 million HES records in 2020 were found to have not met the activity criteria

in the PDS to be included on the 2020 SPD v3 and, of these, around 2m were not included all together and 17m were only active on an income data source. We can unlock many inactive PDS records for inclusion in SPD v4.1 if we assume that we can use HES interactions as a proxy for residence. Now we again have access to income and benefits data, we will be able to confirm the validity of this assumption by comparing interactions across other forms activity.

There exist equivalent datasets that capture Welsh hospital services (Patient Episode Database for Wales). When available for analysis, we intend to incorporate these data into the SPD in a very similar way to what has been done with HES to bring our SPD for Wales in line with England.

Individualised Learner Record

Previous SPD methods have struggled to accurately capture individuals between 16-26. These ages can bridge the gap between school leavers and university inductees/first-time earners and these individuals are therefore prone to being excluded from an admin data approach that requires signs of activity. We identified a lot of records dropping off SPD v3 following a period of ESC enrolment but not then appearing on another source, which impacted the year-on-year consistency and stability of SPD v3 for younger adults (Blake, 2021).

Individualised Learner Record (ILR) is an ongoing collection of data about learners and the learning undertaken by them from providers in the Further Education (FE) and Skills sector in England, underpinning funding and commissioning decisions and the work of Ofsted and other agencies. We identified ILR as a dataset that targets the historically challenging-to-capture age ranges with presence on the ILR in a given year showing an engagement with some FE system.

We were able to match learners on ILR to individuals on the Demographic Index that had been inactive since leaving school as ONS had retained their ESC 'Pupil Matching' reference. Research showed that this ID was congruent between the ESC and ILR. This was only a temporary workaround, and ONS has now linked the ILR into the Demographic Index to make including it in the SPD method more sustainable long-term. Around 430,000 records appeared on ILR in the 2020 reference period but not in SPD v3, primarily aged between 17 and 26.

As with our existing method for the other datasets from the education sector we use, HESA and the School Censuses, we have assumed that presence/enrolment on ILR can be interpreted as usual residence for the 12-month academic year. When we link SPD v4.1 to the Census, we will be able to further test the validity of this assumption by investigating successful links of ILR records specifically.

There exists an equivalent dataset that captures Welsh post-16 learners (Lifelong Learning Wales Record). We intend to incorporate this into our method as soon as we can to bring consistency in our method between England and Wales. ONS is currently in negotiations with Welsh Government to secure this data supply.

Deaths Registrations

Previous versions of the SPD used the date of death information in CIS to highlight and remove individuals in our "spine" that were deceased. Comparisons with the Deaths Registrations dataset indicated that this information flags deaths to a high level of accuracy (Annex 2), but with some areas requiring improvement (Blake, 2020).

Without CIS data in the current SPD v4, there was a risk that the SPD method did not have a research-supported death flag and that larger numbers of deceased individuals were appearing in our population counts.

This was exposed after we incorporated HES activity into the method. We identified that the over-coverage error of our SPD against the mid-year estimates began to increase as we moved up into older ages. We suspected that we may be including some deceased individuals in our SPD due to there being HES activity for them during the 12 months leading up to the reference point.

We linked an early iteration of 2020 SPD v4.1 to Deaths Registrations data covering the 2019 (calendar year) and 2020 (up to 30th June) by NHS Number and around 480,000 records successfully linked. After further analysis, we removed all of these records, the vast majority of which had been included due to HES activity but had a death registration before the mid-year reference date.

One reason that deaths registrations had not been incorporated into the SPD v2/v3 method was the possibility of introducing additional linkage error and removing the wrong individuals. When we are able to link SPD v4.1 to the Census, we will be able to evaluate the commonness of incorrect removals of individuals due to the deaths register, as well as any other coverage challenges associated to deceased individuals in the population.

Inclusion Rules Research

Presence-based PDS rules

As mentioned earlier, the degree of under-coverage present in an activity-based SPD when benefits and income information was removed meant that this method was not fit for use with the DPM. We needed to investigate other ways to exploit more information from the remaining data sources we had access to. It was necessary to establish how far PDS presence alone can be used to infer residence in the population.

We have investigated approaches for identifying and removing records that are in PDS and unlikely to be resident in England and Wales. In pre-processing of the PDS, we remove records without a valid NHS number and those with reason for removal[9] entries (Annex 1).

We also identified a high level of missingness (around 6 million) in the date indicating when an individual last provided their address. Further analysis of these records, such as by looking at their time since last GP registration, suggested they were less likely to still be members of the usually resident population.

Presence-based PDS rules may provide some stability to populations which were previously prone to appearances and drop-offs from SPD v3, such as around life transitions.

Evaluations of SPD v2 suggested that presence on PDS is not a good indicator of continuing usual residence as individuals do not always de-register if they are emigrating or moving away. This results in over-coverage particularly in mobile ages and areas with high population churn and year-on-year list inflation.

Now we are able to reintroduce DWP datasets into the method, we will be able to test assumptions made around the PDS presence against income-related activity. We will be also using address-based data sources such as Council Tax and Electoral Register to validate PDS rules in line with expected household groupings. Additionally, Census linkage will expose areas where the PDS is contributing to over-coverage and signpost other areas of interest where the rules can be tightened.

[9] Details relating to the reason a patient has been (or will be) removed from a Primary Care Provider's list.

Quality Research

SPD v2 and SPD v3 had developed iteratively over a period of years and monitoring of quality throughout the production process of the SPD had not been possible as a result. Throughout the development of SPD v4, we have sought to develop the quality assurance process at all stages of SPD production, from receiving the single sources to producing the final aggregated population stock counts.

We now have a clear set of metrics in place for each source that feeds into the SPD, e.g. missingness, out-of-scope values in key variables, duplication checks and linkage rates to the DI, which have provided extra confidence in the figures we produce.

For SPD v4, the pre-processing checks have been:

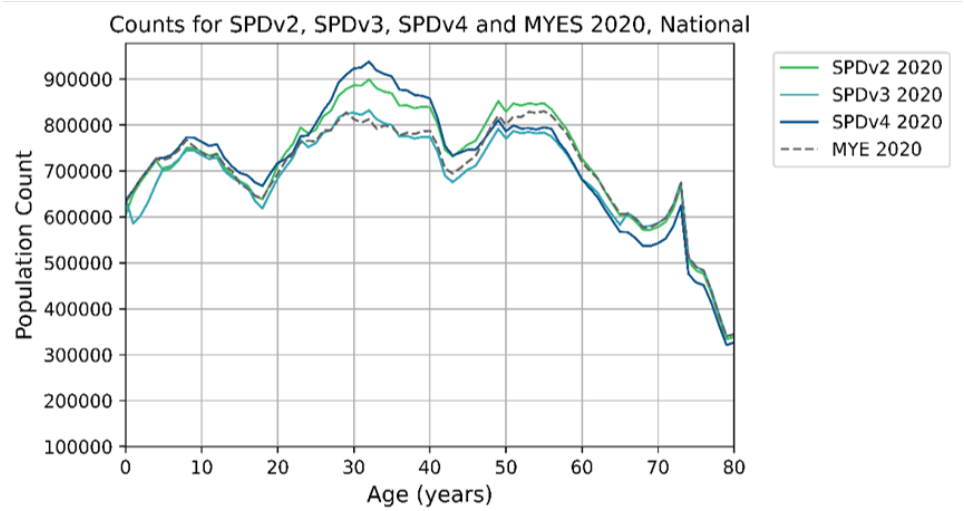
1. Standardised, so that the same checks are conducted at each stage for each source.
2. Developed, using feedback from the Methodology Division Quality Assurance report on SPD v3, research conducted by Statistics New Zealand, and wider research across the transformation projects into best practice on quality.
3. Automated so that they are produced automatically during the processing.

This has meant that we can better assess whether the process introduces/reduces/increases bias or error, assess the linkage methods and assess the quality of the inclusion rules and method of assigning attributes. We are now seeking to do similar development for the other stages of the SPD build after the pre-processing checks outlined above in 'How is an SPD made?' including filtering rules and assigning attributes.

Alongside our research we are streamlining the production of SPD v4 into a Reproducible Analytical Pipeline (RAP). The purpose of this is to ensure that our processing is efficient, allows the production of quality metrics throughout the process, and that the method is auditable. It also means that as we continue to develop the SPDs we can plug in new elements to assess their impact, whilst easily being able to revert to an earlier version.

Analysis of SPD v4.1

Returning to a largely presence-based system with the PDS increases the possibility of observing some year-on-year list inflation in the SPD. We have made some aggregate-level comparisons across the 2016-2021 time series of SPDs built to the v4.1 method to understand the levels of list inflation of cohorts and record-level appearances.



We have compared the 2020 England and Wales population totals by age across the different SPD versions, and with the official mid-year estimates. Our analysis shows using PDS presence alone results in higher coverage for most of the younger population (up to approximately age 40) and lower coverage for the older ages. The higher coverage observed in the 20-40 age group is mostly driven by males on the PDS.

Development of SPD v4.2

Producing SPD v4.1 has allowed us to develop and test our processes and show what is possible without DWP data. The results show how challenging it is to produce estimates with reasonable coverage, without regular activity data, particularly income data. Our progress on SPD v3 shows just how valuable a contribution this data makes.

We have integrated the CIS and BIDs data into SPD v4.2. The income data we now have access to is a little different to the data we have used in our SPD methods previously, and there are some new data sources. To assess the impact of using these data sources, we have started by applying the same income activity rules as SPDv3. Our aim is to directly compare the 2020 SPDv3 with the 2020 SPDv4.2 to assess the impact of the changes in the data and the new sources introduced in SPDv4.1. We will then look to adjust the income activity rules if necessary to improve coverage.

We are also considering the feasibility of using HMRC PAYE-RTI data in place of, or alongside, our existing annual PAYE summary dataset, to further improve the income activity rules. This is a large, complex dataset which is likely to take some time to fully understand.

For now, we have removed the PDS presence rule as we explore the feasibility of using address-based intelligence from Electoral Register and Council Tax, to improve this further. We will then assess whether this rule is likely to improve the coverage of the income-based SPDv4.2, particularly for those people who do not have signs of activity in the datasets we use.

To summarise, for SPDv4.2, we are using the following data sources;

- Patient Demographics Service (PDS) – *National Health Service*
- Customer Information System (CIS) – *Department for Work and Pensions*
- Higher Education Statistics Agency (HESA)
- English School Census (ESC) – *Department for Education*

- Welsh School Census (WSC) – *Welsh Government*
- Benefits and Income Dataset (BIDS) – *Department for Work and Pensions (with HMRC PAYE data)*
- Deaths Registrations
- Hospital Episode Statistics and Emergency Care Dataset (HES) – National Health Service
- Individualised Learner Record (ILR)- DfE

We have applied a combination of SPDv3 activity-based rules, with SPDv4.1 activity rules for HES, ILR and Deaths Registrations.

We will also keep up to date with the landscape of data available across the office and identify new sources for our method. Once data delivery is finalised, we expect 'quick wins' with:

- **Patient Episode Database for Wales** – bringing hospital services activity in the SPD in line between England and Wales
- **Lifelong Learning Wales Record** – bringing coverage for post-16 learners in line between England and Wales

Over the coming weeks we will be evaluating the quality of this SPD v4.2 by comparing it to previous versions and using the Census/Census Coverage Survey (CCS) to DI linkage analysis. Our analysis will help us confirm the combination of data sources and rules for our best SPDv4. We hope to share some results at the November session.

Census/CCS to DI linkage

We are undertaking a piece of analysis to explore record-level linkage between the Census/Census Coverage Survey (CCS) dataset and Demographic Index (DI). This will provide a rich evidence base to inform decisions for the statistical transformation of population and migration outputs, including the DI, Dynamic Population Model (DPM) and the SPD.

Data linkage was conducted between the DI and the 2021 Census to Census Coverage Survey linked dataset (CC), which had been linked together in a previous linkage project. The linkage project included using automatic linkage methods on the whole of the CC to the DI, and then using clerical matching to ensure the highest quality was obtained. Due to resource limitations, only a sample known as 'CCS2' of postcodes was linked in full (using both automatic and clerical matching) to the quality required for this analysis. Further detail on this sample of CCS2 postcodes is noted under assumptions. A quality analysis to estimate linkage accuracy within the CCS2 was conducted and suggests that the DI-CC linkage has a precision of between 99.3% and 99.7%, and recall between 99.1% and 99.7%.

An observation found, when using clerical resolution methods between the CC and the DI, was that the quality of the administrative data and reduced number of variables in comparison to CC did make clerical matching decisions harder. This causes ambiguity in linkage accuracy and bias where we are unable to confidently identify errors and hinders our ability to link. Even with this observation, the quality of the linkage between the CC and the DI in CCS2 areas is still considered to be good. This analysis has been recently reviewed by Methods and Research Assurance Group (MaRAG), Longitudinal Study Advisory Panel (LSAP) and Methodological Assurance Review Panel (MARP) and began in September 2022. Since the work presented in the paper 'Linkage of the

Demographic Index to 2021 Census Coverage Survey' (EAP177) we have spent time investigating some of our assumptions and planned methods.

Assumption 1: Using 2020 HESA will be sufficient

Due to the time constraints of the project, the 2020 extract of Higher Education Statistics Agency (HESA) data was used during the DI-CC linkage. To understand the potential impact of using HESA 2020 data instead of HESA 2021 data, analysis was carried out exploring the quality of address information in both extracts. 2020 HESA was linked to and compared against 2021 HESA to assess: how many of those found in HESA 2021 will be missing in the DI-CC linkage analysis; how many students left in HESA 2020 would be found in the linked data and what the geography of those on both extracts looked like. 2020 HESA was also compared to 2021 PDS data to assess the accuracy of the geography of student leavers (e.g. how many students update their address on administrative data within a year of finishing university?). This can help us to understand how many people are being incorrectly included in analyses on students, as well as how they might act in administrative data, providing context to conclusions drawn for these populations.

Initial analysis found that, using 2020 HESA instead of 2021 HESA will not affect the DI-CC linkage analysis greatly, as less than half of people who appear on 2020 HESA, but not on 2021, updated their address on the PDS within a year of leaving university. However, there are some exceptions. When conducting analysis on 18-20 year olds, it should be noted that there are many who appear on 2021 HESA, but not on 2020 HESA, and will therefore be missed in our analysis. Consequently, they may have less accurate geography within their ONS ID cluster, as they are missing from an additional administrative data source. When conducting analysis on 22-25 year olds, it should be noted that there are many who appear on 2020 HESA, but not on 2021 HESA, so are being incorrectly captured in analyses involving student populations. We could therefore see records being incorrectly matched on geography data or having presence in the DI where they should not. We are reviewing how we can integrate HESA 2021 using the latest DI build, to further reduce the implications of the 2020 HESA being used. Further consideration to the use of 2020 HESA will be given as we start to produce analyses.

Assumption 2. Using CIS 2020 will be sufficient

It is worth noting that while the DI was constructed using Customer Information System (CIS) data from 2020, CIS data from 2021 is being linked to the 2021 Census and a look up between the CIS 2020 and CIS 2021 data is being used to bring them together. The CIS data is hashed, so the Census linkage variables were hashed and linked to CIS in a separate space and method to the other sources in the DI, and brought together later.

The CIS data was resupplied towards the start of the project, meaning the CIS data was removed and completely resupplied, so the 2021 rebulk contained the 2020 data, although further cleaning and edits may have been made to the data before reaching ONS. We have reviewed the implication of how the newly supplied CIS data may have changed the construction of the ONS IDs within the DI (i.e. new information or better quality cleaning could lead to ONS IDs splitting or merging) and we found less than 0.05% of CIS master keys linked to a different ONS ID in the DI2.1 than in DI2.0.

Assumption 3. Effects of coronavirus (COVID-19)

A key consideration ahead of the analysis of the DI-CC linkage was how the coronavirus pandemic may have impacted the quality of both the administrative data sources within the DI as well as the Census and CCS, in addition to the quality of the linkage itself. The widespread displacement of

people throughout the pandemic period could have led to inaccurate address information on both administrative and Census data. International migrants who returned to their country of origin during the pandemic, when they otherwise would have remained in England and Wales, may not have a Census return and thus appear as DI over-coverage. There is also some concern that the Census response for Special Population Groups, such as communal establishments, was impacted by coronavirus. Linkage of Census/CCS to the DI could therefore compound known issues in the administrative data in accurately capturing these populations.

It should be noted, however, that Census Field Operations suggested that there is little evidence to indicate that coronavirus caused issues with household response rates. Indeed, the lockdown restrictions in place at the time of collection are thought to have contributed to the 97% response rate (ONS 2022) that was achieved. Furthermore, it is likely that the national vaccine rollout acted as a prompt for people to update their administrative health records, a theory supported by preliminary analysis of GP registrations throughout that period. While more accurate information is beneficial for the linkage process, this raises the question of the applicability of findings from this research to the post-pandemic picture, given data were collected primarily in 2021. Consideration will therefore need to be given to how we use the results to feed into any wider decision-making.

Assumption 4. Moving from CCS areas to whole population

Note that due to time and cost restraints, clerical work could not be completed for the entire Census/CCS and DI. We therefore restricted clerical work to just those DI or Census/CCS records that have a postcode in a CCS area. This is defined as when either an ONS ID contains one source in a CCS area or the Census/CCS usual or alternative address is in a CCS postcode.

The amount of clerical work was still too great to fit into the given time-scale. The clerical was therefore restricted further to a subsample of approximately half the CCS postcodes. This subsample was selected by including CCS postcodes as follows:

- If there is only one CCS postcode in an output area, then select it
- If there are two or three CCS postcodes in an output area, then randomly select one of them
- If there are four or five CCS postcodes in an output area, then randomly select two of them

This method of selecting the subsample should mean that the subsample is stratified in the same way as the original CCS sample and will include postcodes from all over England and Wales.

This concept was agreed at MaRAG so has been discussed at assurance groups but also with the linkage and estimation expert groups. We are currently reviewing weighting options.

Comparing the coverage of the SPD, DI and Census

The project includes three questions specifically targeting SPD v4 development. We plan to code these analyses in a reproducible way, allowing us to run the same evaluations with SPD v3 and SPD v4.2. There is further research related to geography, communal establishments and coverage of the DI, which will also inform SPD development.

This analysis will help us identify how well our inclusion rules are working for different groups of the population and how well we can allocate people to their usual residence using admin data. We will be exploring both under-coverage and over-coverage, as well as which administrative data sources are most likely to provide us with a usual residence address. We will be exploring how this differs for different sub-groups of the population.

Future Development and Next Steps

Delivering our “best-possible” SPD version 4 for the Recommendation

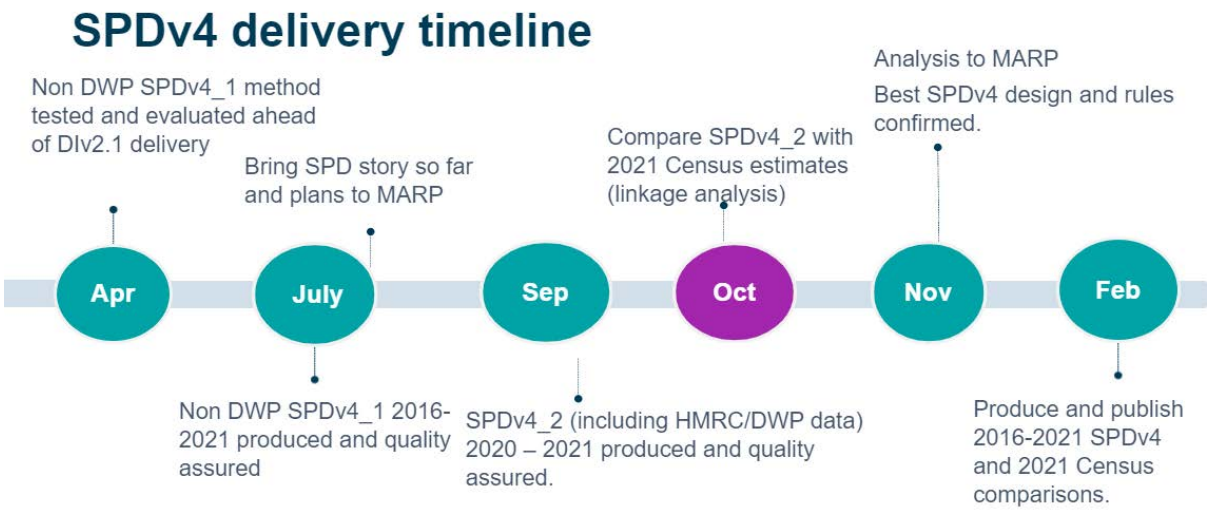


Figure 1: SPD v4 development 2022 roadmap

Producing SPD v4.1 allowed us to develop and test our processes and show what was possible without DWP data. The results showed how challenging it was to minimise over-estimation without regular activity data, particularly income data. Developing SPD v4.2 has allowed us to integrate DWP data, bringing us closer to producing a “best possible” SPD for 2021. Figure 1 shows during the remainder of 2022, we intend to evaluate the quality of SPDv4.2 using the Census/CCS to DI linkage analysis and use this analysis to help us confirm the combination of data sources and rules for our best SPDv4. This comparison will feed into the NS Recommendation and inform future development of the SPD.

Longer term developments

We expect the results of the Census/CCS-DI linkage analyses to provide clear evidence for specific developments required in the SPD method as we begin to look beyond SPD v4.2. Improvements will also be made towards the coverage and quality of the underlying Demographic Index as a result of other research questions in the project and we expect these to feed through into our SPD as well.

We will be exploring how timely we can make the SPDs as our data feeds and our methods mature. Currently the HESA data we use isn't available until February following the reference date and this will be the constraining factor for a final SPD. We will explore the potential for a provisional and final SPD to bring greater timeliness to the initial feed into the DPM. This will be dependent upon evaluating the quality of this approach and whether it provides additional benefit to the DPM outputs.

Modelled inclusion rules

We have begun exploring the potential applications of machine learning to address some of the challenges associated with producing an SPD. Generally, supervised classification may have

applications where our use of data requires decisions to be made around how it is processed, integrated or labelled or where the probability of these decisions being made correctly needs to be predicted. Specific applications include training data-driven models to:

- Classify/fractionally weight admin records as describing members of the target population instead of using filter rules;
- Selecting/fractionally weight attribute values (e.g. address) rather than selection rules;
- Classify/fractionally weight data linkage as true/false matches;
- Assign categorical attribute values where they are missing in admin data.

Estimation

For the DPM to produce reliable and unbiased population estimates, it requires the SPDs to be unbiased. Coverage errors observed in the SPDs need to be corrected by using estimation methods to calculate appropriate factors that can be applied to the SPD. The paper 'SPD Estimation Options' (EAP184) discusses potential approaches for achieving this aim along with their benefits and limitations.

Determining when to stop/start SPD development in future

Above we have outlined how the SPD has evolved over a period of several years. The evolution of the SPD has been influenced by both the available admin data and the wider system that it will form a part of. For example, the design of SPDv3 to have deliberate undercoverage that might facilitate the use of a dual-system estimation (DSE) approach to estimate the population.

The role of the SPD as envisaged now will primarily be as an important cross-sectional stock estimate of the population to feed into the Dynamic Population Model (DPM). In this context a stable SPD design with associated quality metrics is needed quickly so that it can be used in evaluating DPM options and performance ahead of the 2023 recommendation.

However, it is important to have a set of criteria against which we can decide whether the SPD has reached a point where we can agree that the design should be stabilised, but also gives the flexibility to decide in future to revisit the research should things change. These relate to the National Statistics Code of Practice principles of trustworthiness, quality and value.

1) The design should be reliable and reproducible and be capable of clear articulation to users.
2) The purpose of the SPD should be clearly articulated – for example, is it only to provide aggregate level inputs to the DPM, or does it have value as an output in its own right.
3) The accuracy (both variance and bias) of the estimates that can be measured and produced using the SPD once an appropriate estimation system has been developed.
4) The variance and bias can be measured independent of a census and should be capable of being measured on a regular and ongoing basis.
5) The variance and bias are within tolerance levels to be set, whereby if the error was to exceed these levels we should restart development to look for improvements.
6) Where significant changes are planned or observed in the key administrative sources or to any survey that might be used as part of the estimation process, research should be established to determine the likely impact and assess how to further develop the SPD, the estimation system or both.
7) Where new administrative sources become available that have potential to improve the SPD significantly, research should initially be undertaken to assess this potential.

In all of the cases above there will need to be cost/benefit trade off in any decision to re-open the research. This should also take into account any imbalance in the way the improvements might

impact across age/sex/geography and therefore have the potential to impact the allocation of resources.

Conclusion

SPD v4 is still in its development phase and we will continue to develop a second iteration in the areas and to the timelines outlined above to ensure we deliver valuable SPD information to the Dynamic Population Model. The value of this will depend on developing a robust, intuitive and research-led method which has been extensively evaluated against official estimates, informing our understanding of its quality and defining the estimation challenge.

We have been able to extensively research how far we can approximate the usual resident population from admin data without income and benefits information. The ongoing development of version 4 will be dependent on reintroducing DWP sources into our method to validate recent decisions taken with greater evidence. Linking our SPDs at a record level to the Census estimates will provide additional rich insight to test these rules and establish the next steps we need to take in this journey.

The progress we can make across the Office towards coherent admin-based population and migration estimates will move us towards meeting the needs of our users for more timely and accurate insight and will be central to the 2023 Recommendation. Longer term, our SPD development will be customer focused and will require further user engagement to identify how and where we should target improvements.

References

Bartleet, L., 2021. *2023 Recommendation Guiding Aims.* [Online]
Available at: https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2021/11/EAP157-2023-Recommendation-Guiding-Aims.pdf

Blackwell, L., 2020. *Indicative uncertainty intervals for the admin-based population estimates: July 2020.* [Online]
Available at:
https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/indicativeuncertaintyintervalsfortheadminbasedpopulationestimatesjuly2020
[Accessed 27 July 2020].

Blackwell, L., 2022. *Dynamic population model for England and Wales: July 2022.* [Online]
Available at:
https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/dynamicpopulationmodelforenglandandwales/2022-07-14
[Accessed 14 July 2022].

Blackwell, L., Elliott, D. & Bryant, J., 2021. *Integrated statistical design for the transformed population and social statistics system - Bayesian methods for demographic estimation.* [Online]
Available at: https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2022/03/EAP174-Bayesian-Methods-for-Demographic-Estimation.pdf

Blake, A., 2020. *Measuring and adjusting for coverage patterns in the admin-based population estimates, England and Wales: 2011.* [Online]
Available at:
https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/measuringandadjustingforcoveragepatternsintheadminbasedpopulationestimatesenglandandwales/2011#reasons-for-overcoverage

Blake, A., 2021. *Developing admin-based population estimates, England and Wales: 2016 to 2020.* [Online]
Available at:
https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/developingadminbasedpopulationestimatesenglandandwales/2016to2020

Office for National Statistics, 2013. *Beyond 2011: Options Report 2.* [Online]
Available at:
https://webarchive.nationalarchives.gov.uk/ukgwa/20160108193314mp_/http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/beyond-2011-options-report-2--o2-.pdf

Office for National Statistics, 2015. *ONS Census Transformation Programme.* [Online]
Available at:
https://webarchive.nationalarchives.gov.uk/ukgwa/20160105160709/http://www.ons.gov.uk/ons/guide-method/census/2021-census/progress-and-development/research-projects/beyond-2011-research-and-design/research-outputs/administrative-data-research-outputs--201

Office for National Statistics, 2016. *Methodology of Statistical Population Dataset V2.0.* [Online]
Available at:
https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/methodology/methodologyofstatisticalpopulationdatasetv20

Office for National Statistics, 2017. *Research Outputs: Estimating the size of the population in England and Wales, 2017 release.* [Online]
Available at:
https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusresearchoutputs/sizeofthepopulation/researchoutputsestimatingthesizeofthepopulationinenglandandwales2017release#how-do-the-2016-estimates-p

Office for National Statistics, 2019. *Developing our approach for producing admin-based population estimates, England and Wales: 2011 and 2016.* [Online]
Available at:
https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/developingourapproachforproducingadminbasedpopulationestimatesenglandandwales2011and2016/2019-06-21

Annex

Annex 1: Data cleaning and pre-processing applied to data sources in SPD version 4.1

All of the sources use date of birth to establish an age on the mid-year reference point and have geography files joined to them to place records in up-to-date Local Authorities.

The filtering and deduplication processes in the method differs by source:

Source	**Filter to records:**	**Deduplicated?**
PDS	• with GUID (ONS identifier) • without Reason For Removal date • with address in England / Wales • with Reason For Removal type	Yes
ESC / WSC / HESA	• active in the academic year • with GUID (ONS identifier) • with addresses in England/Wales	No
HES	• with a valid NHS number • not requesting anonymisation • with birthdate after 1900 • with a birthdate before the reference date • with male/female sex • with postcode • active with services in the 12 months prior to reference date year • without a death flag • with address in England / Wales	Yes
ILR	• in the academic year • with address in England/Wales • showing activity through learning record	Yes

Annex 2: Comparison of deaths registration data to SPD v2 and CIS

Comparisons to SPD v2

To enable comparison with the 12 months prior to the SPD reference date, July to June each year, the deaths register has been split into quarters by date of death. Therefore, deaths occurring in quarters 3 and 4 (1st July to 31st December) of one year and those occurring in quarters 1 and 2 (1st January to 30th June) of the following year have been combined to match the yearly coverage of the SPD. In doing this some late registrations whose date of death did not occur in the year they are registered are missed, as there is no way of telling which quarter they were registered as a registration date is not contained in the data.

There are some records on the Death Register for people whose place of usual residence is Scotland, Northern Ireland or 'other'. These were removed as we are only interested in the deaths of those who are usually resident in England and Wales. The deaths register was then linked to the SPD by NHS number.

The research looked at how many deaths were still included in SPD v2 that should have been removed in the SPD v2 2015 and 2016. All deaths occurring and registered between the 1st July 2014 and the 30th June 2015 were linked to the SPD v2 to see if there were any deaths remaining on the SPD that should have been removed. The same was done for 2015 to 2016 deaths which were linked to SPD V2 2016 (Table 1). The majority of deaths occurring during this period were not found on the SPD confirming they had been correctly removed.

Table 1: Death register linked to the SPD v2, England and Wales, 2015 to 2016.

	SPD v2 2015	**SPD v3 2016**
Death registrations found on the SPD	1,008	1,097
Death registrations not found on the SPD	507,229	495,168

Table 1 also shows that just over 1,000 death records were not removed from the SPD V2.0 estimates in 2015 and 2016. Most of these are deaths that occurred at the end of the reference period. Although this is a small number of records, they represent deaths still being counted as part of the SPD population estimates.

Comparisons to CIS

The deaths registrations data were linked to the Customer Information System (CIS) 2016 to determine if dates of death are consistent across the two administrative data sources (Table 8). Most of dates of death were consistent across both data sources (97.74%). Only small differences were found where the dates of death did not match across the administrative data sources. There were nearly 1,200 deaths where the CIS date of death was earlier than the official date of death which would suggest an error in the CIS data with the accuracy of the death register being accepted as the truth. Conversely, around 670 deaths had an earlier date of death on the registrations data when compared with the CIS. Again, the register entry is considered to be correct in view of the quality checks carried out at the time of the registration of the death.

Table 2: Comparison of dates of death on the deaths register and the CIS for 2016

	Same date of death	Death registration date of death earlier than CIS entry	CIS date of death earlier than death registration entry	No date of death on CIS
Count	478,924	666	1,199	9,161